Geo-referenced data and networks for measuring risk

March, 18, 2022

Gian Paolo Clemente
Università Cattolica del Sacro Cuore, Milano
gianpaolo.clemente@unicatt.it

Francesco Della Corte
Università degli Studi di Roma, La Sapienza
francesco.dellacorte@uniroma1.it

Diego Zappa
Università Cattolica del Sacro Cuore, Milano
diego.zappa@unicatt.it

Aim of this work

- Aim of the paper is to show how the spatial objects and the information concerning the structure of the roads, that can be collected from open data sources, together with the crash history can be used to map the risk related to each road.

- To achieve our aim, we need to adapt the current methodology about geospatial modelling to the constraints derived from the maps of the roads of a particular area and to exploit supervised/unsupervised statistical learning algorithms to estimate the local risk of frequency (and severity).

- We follow a combined approach:
  - A statistical model will be developed in order to assess the risk on the basis of a set of features related to the characteristics of the streets.
  - From the spatial object we build a weighted network, where vertices and edges correspond to geographical elements as junctions and roads and where the assessed risk of each segment is used as a weight.

- The research can be classified within the big data paradigm not only because the data used will be really "big" but, most of all, because it needs to merge and exploit information coming from different and, in some cases, unstructured sources.

Aim of this work

Most of the papers published in the literature look for an “ex post” spatial modelling of crash risk: starting from the company database and the knowledge of accidents’ location, they estimate the spatial dependence of the risk (see, e.g., Assunção et al. 2014).

We want to contribute to the literature proposing a different approach to measure the risk of collision. By extensively using georeferenced applications and information regarding the traffic and road characteristics (length, number of crosses, highways, etc.) that can be obtained by standard accessible web applications, it is possible to make further inference on the collision risk of where a driver usually drives his/her car with limited costs.

- In particular we focus on “where the policyholder drives”

We do not consider here (actually a research in progress) other features that can be detected by telematic data:

- Driving behaviour (see, e.g., Wuthrich, Buser, 2019)
- Data about driving habits (e.g. KM, daytime, weather conditions, etc.)

The data

We make use of some specific data and we will focus in the application on Milan area (city and province)

1. **Road Network and its characteristics**: data can be downloaded freely from the web to create the street network. For each segment of the network many information are freely available, that can be used as covariates to explain spatial point objects (e.g. location of accidents)
2. **Location of risk accidents:** Using open source datasets provided by the Italian national office of statistics (ISTAT), which records information on the location of all car accidents that resulted in fatalities or injuries of at least one person, we had the possibility to project crashes on the road network of Italy. The dataset can be augmented using local socio-demographic features (e.g. population density, concentration of families, housings etc.).
3. **Knowledge of the trajectory covered by the drivers:** this is the most critical information. At present we have the availability of data saved through thousands of black boxes (but detailed results will be masked because of data confidentiality)

The data (road characteristics – an example)

id_link	highway (type)	highway (length)	URBAN	# junctions	# pedestrian crossings	# traffic lights	# car crashes
1	Tertiary	119	Y	5	0	0	0
2	Secondary	309	Y	5	2	0	0
3	Primary	11.3	N	4	0	3	0
4	Primary	11.3	Y	5	1	0	2
5	Primary	150	Y	7	2	1	0
6	Secondary	35.4	N	6	0	0	1
7	Secondary	67.9	Y	6	0	3	0
8	Tertiary	97.7	Y	6	1	0	3
9	Motorway	157	N	4	0	0	0
10	Other	150	N	6	0	1	1

For each OSM segment save/compute

- Type of road (highway)
- Features (if available) e.g. surface, maxspeed, lit…
- Number of junctions (computed exogenously: proxy very close to reality)
- Number of traffic lights
- Number of pedestrian crossings

Some details

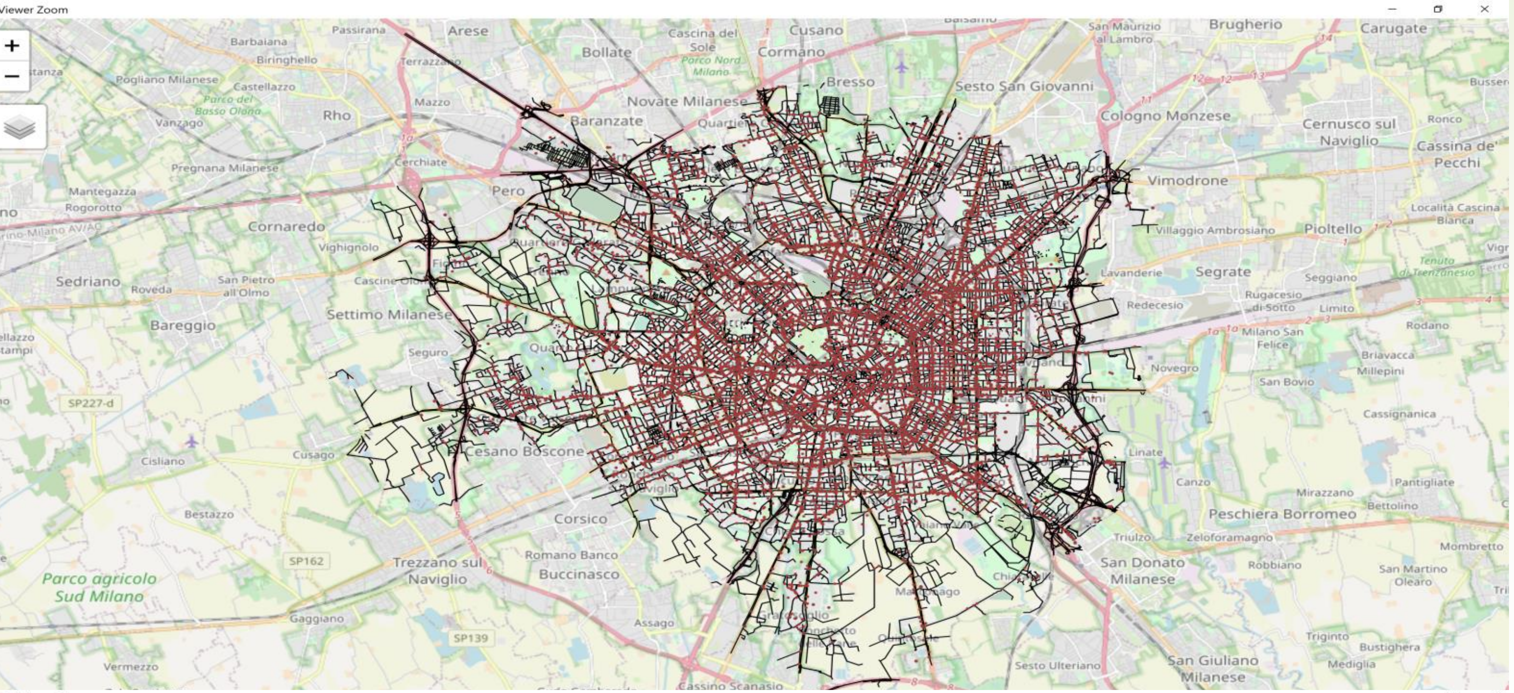
A bit of bias in the data is somehow unavoidable in this type of research:

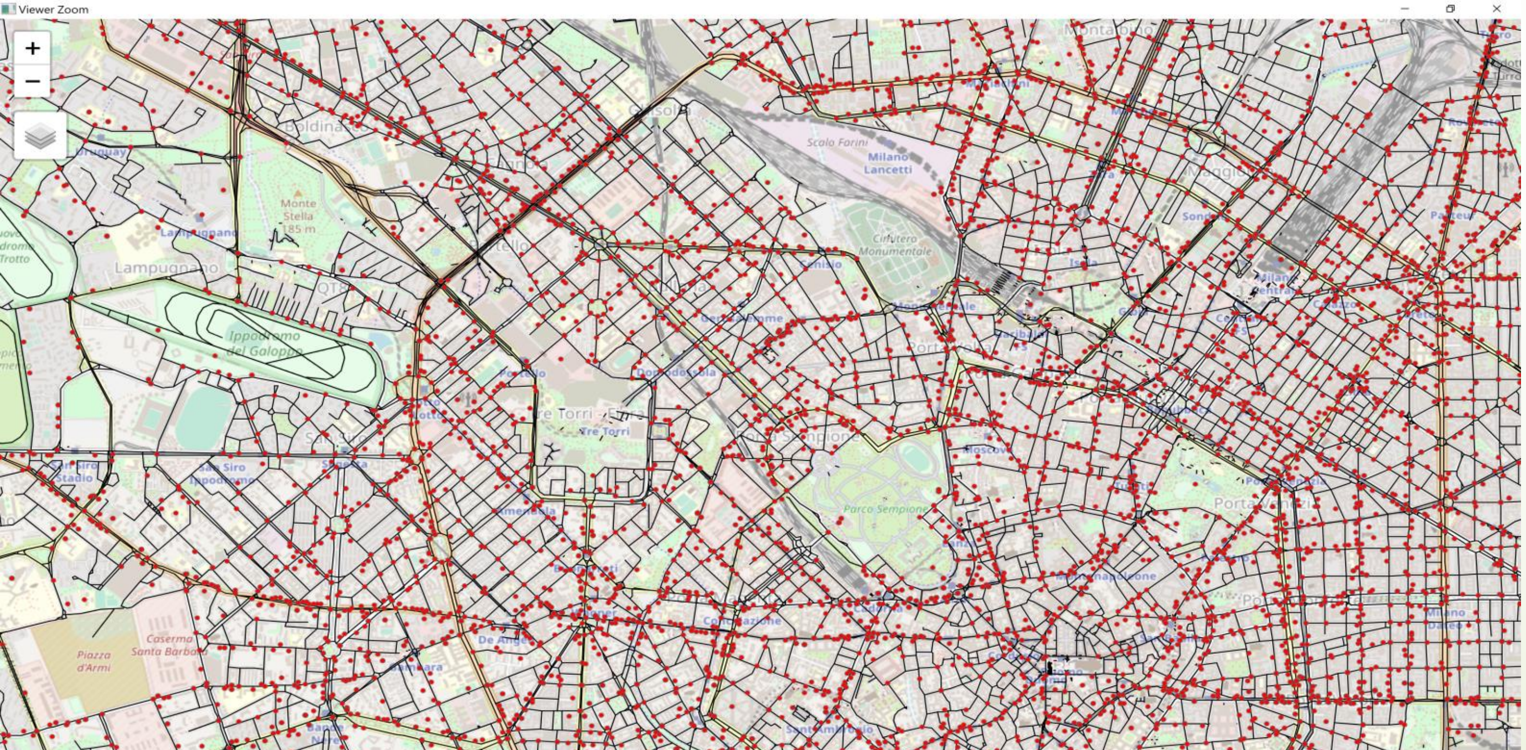
1. The number of road crossings are not directly available in the OSM database.

   For each road, we computed it as the number of segments that have in common one coordinate with that road. This method represents an approximation of the true crossings (consider, for instance, two roads at different level one above the other through a bridge) but it returns in general an estimate quite close to reality.

2. Coordinates of accidents are not always strictly in line with a segment. Approximations are due to proxies implicit into the reverse geocoding algorithms or to errors in the registration of accident locations. We project (orthogonally) that coordinates onto the closest segment

3. Restricted to some regions, many features in OSM database are not available and thus not useful for model identification.





An example of roads and accidents

- We focus on Milan area and we display the road-map and the accident locations (in red)
- On the bottom, a zoom on a central area of Milan

The Model

1. To estimate the risk of accident at segment level a global model cannot be fitted. At least a mild but not zero correlation must be locally considered.

2. **To include that we have split the domain into subregions.** We report the results obtained considering sub-areas based on ZIP codes (other tessellation criteria are under evaluation)

3. We follow the following steps for each area:
   - Select the data (claims and road characteristics) of a specific subregion for a specific year.
   - Enlarge the area including a buffer on the borders
   - To consider spatial dependence, for each segment and for each covariate, add new covariates based on the combination of the characteristics observed for that covariates in the surrounding area and weighted on the inverse of the distance (see next slides for details). Only segments within a selected buffer are considered.
   - Fit a glmnet model considering alternatively Poisson and Negative Binomial distributions and vehicle miles travelled (VMT) measure as offset.
   - Estimate a risk measure for each segment
   - Smooth locally the results to reduce anomalous peaks if present.

4. It is noteworthy that:
   - on each area a different model is customized.
   - Other criteria for tessellations are, at moment, under evaluation (e.g. administrative boundaries, Voronoi tessellation) and the best criterion will be further identified.

Network-based distances

- To compute distances, we convert the street network in a *graph*, i.e., a mathematical representation consisting of nodes connected by edges (or arcs in the directed case) loaded with weights (or labels), that can be directed or undirected.
- In particular, **we focus on a "junction graph"** (see, e.g., Marshall et al., 2018), where each segment is an arc and nodes are given by junctions (or by termination of closed streets).
- Formally, given the street network, we build a graph $G \;=\; (V; E)$ where V and E are respectively the set of n vertices and m arcs. Two nodes are adjacent if there is an arc $(i, j) \in E$ (i.e. a road segment) connecting them
- If a weight $w_{i,j} > 0$ is associated with an arc (i, j) , a weighted graph $G \;=\; (V; E; W)$ is obtained, being W the set of weights.

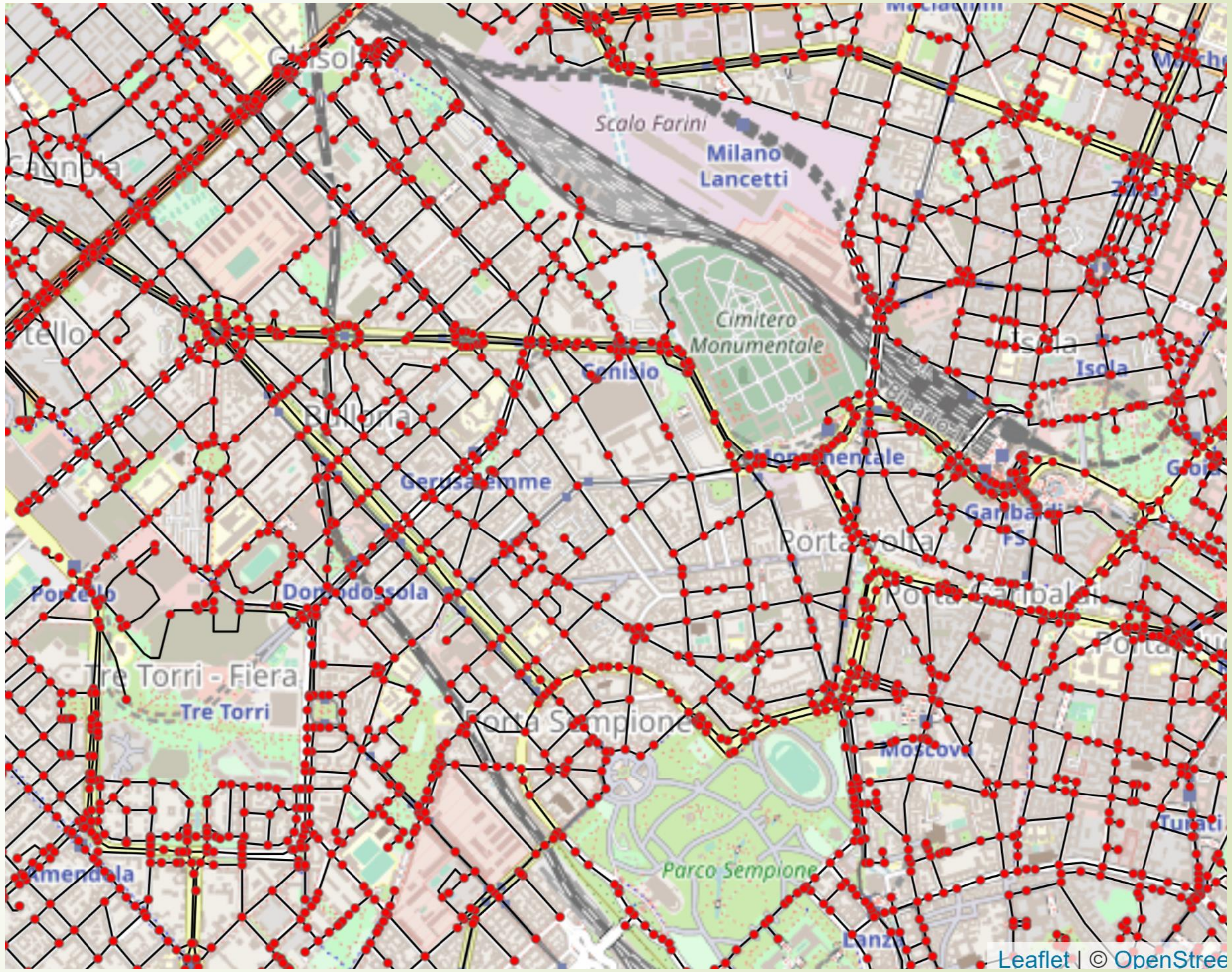
In general, both adjacency relationships between vertices of G and weights on the arcs are described by a nonnegative, real n-square matrix $\boldsymbol{W}$. In the unweighted case, this matrix becomes the classical binary adjacency matrix $\boldsymbol{A}$, of entries $a_{i,j}$, where $a_{i,j} = 1$ if $(i;\; j) \in E$ and 0 otherwise.

Network-based distances

- In particular, we consider at moment ***a directed and weighted network*** $G_w = (V; E; W)$ equal to G, where each arc *is weighted with the length of the segment.*

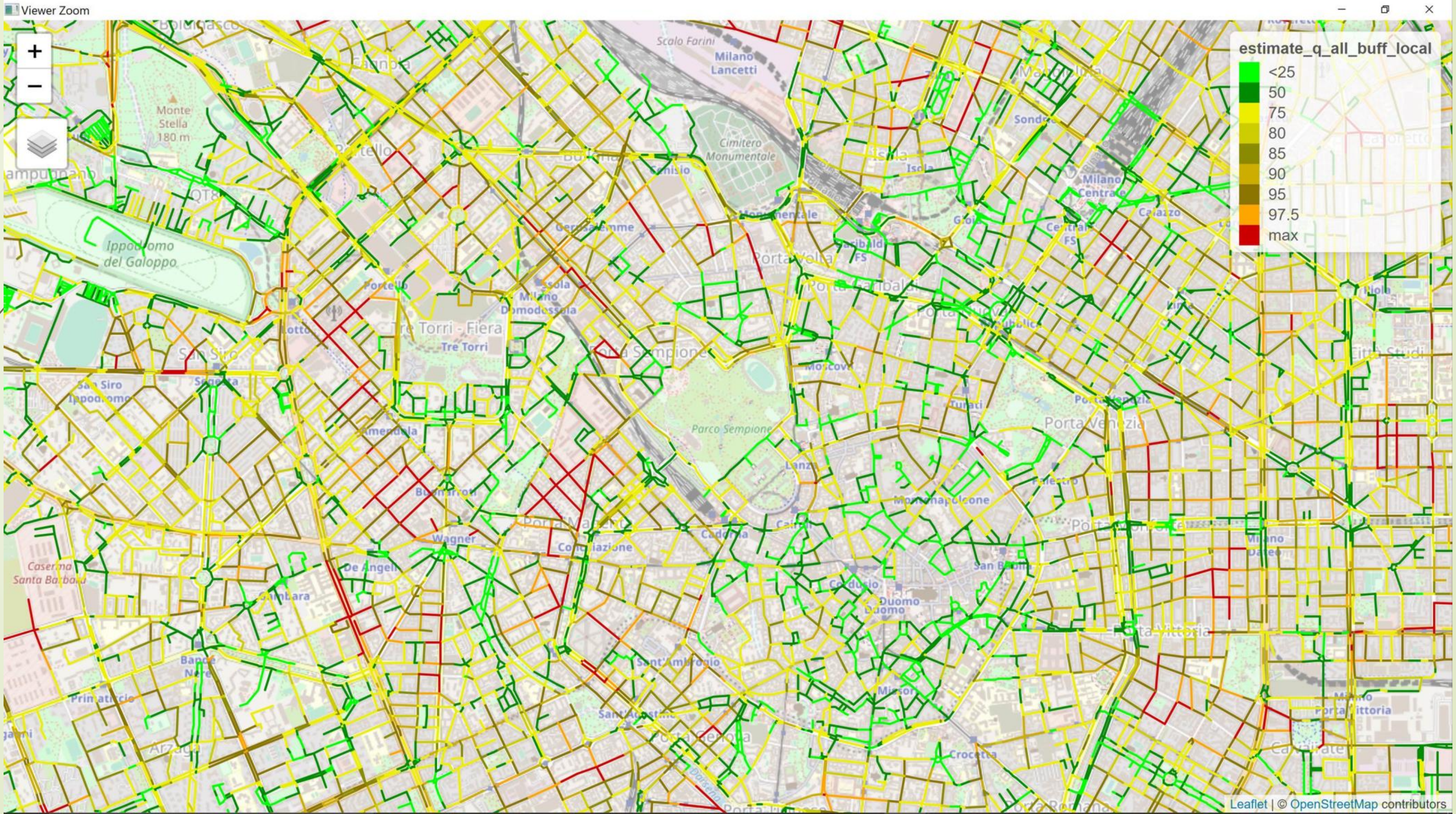
- Distances between two roads have been computed by adding centroid to each segment and by computing the directed weighted shortest path between two centroids.

- The **shortest path problem** is the problem of finding a path between two nodes in a graph such that the sum of the weights of its constituent edges is minimized.

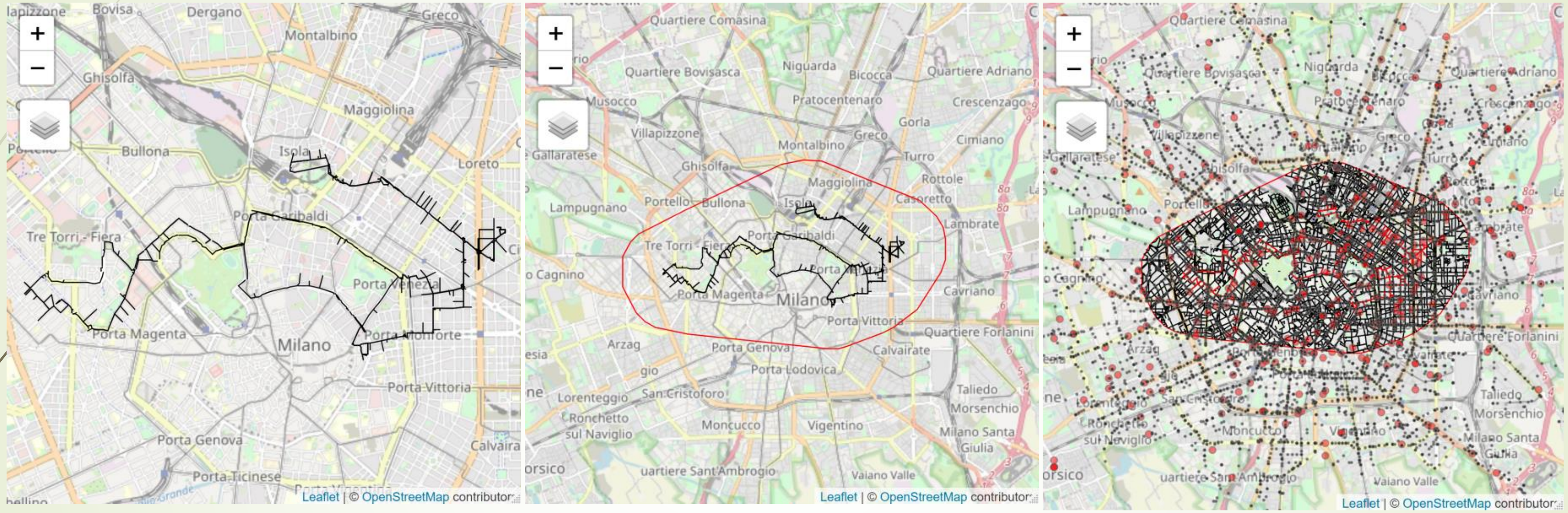


First Results (map of the risk)

Map of a central area of Milan (close to the main center)



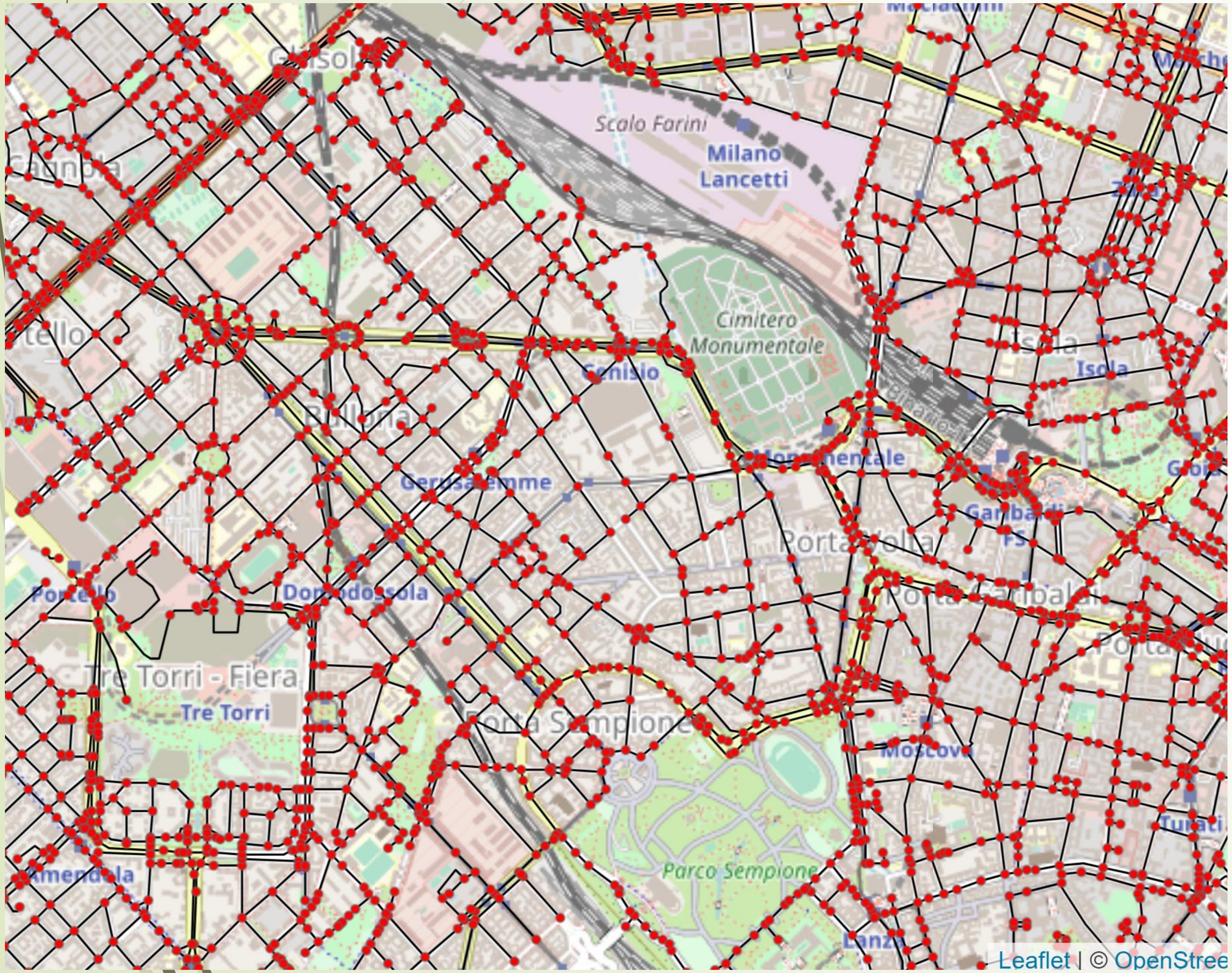
Analysis for each device



By means of the risk previously assessed, we can estimate the risk to each trip of a device and the total year-risk of the device.

Alternatively, it is possibile to assess the risk related to the surrounding area of the trip building a proper envelope.

Some additional analysis based on networks



- We deal now with two types of network:
  - $G = (V; E)$ ***an unweighted network*** with n nodes (junctions/road terminations) and m arcs (road segments)
  - $G_w = (V; E; W)$ ***a weighted network*** equal to G, where each arc ***is weighted according to the risk of the segment*** detected at previous step.
  - Considering only Milan and province, we have an unweighted network with the following characteristics:
  - 142,497 nodes, 171,079 edges.
  - The network is very sparse (density is close to zero)
  - The assortativity and transitivity are also very low (both around 0.03)

Topological Indicator: betweenness

- We focus here on the topology of the network, assessing the global importance of network elements.

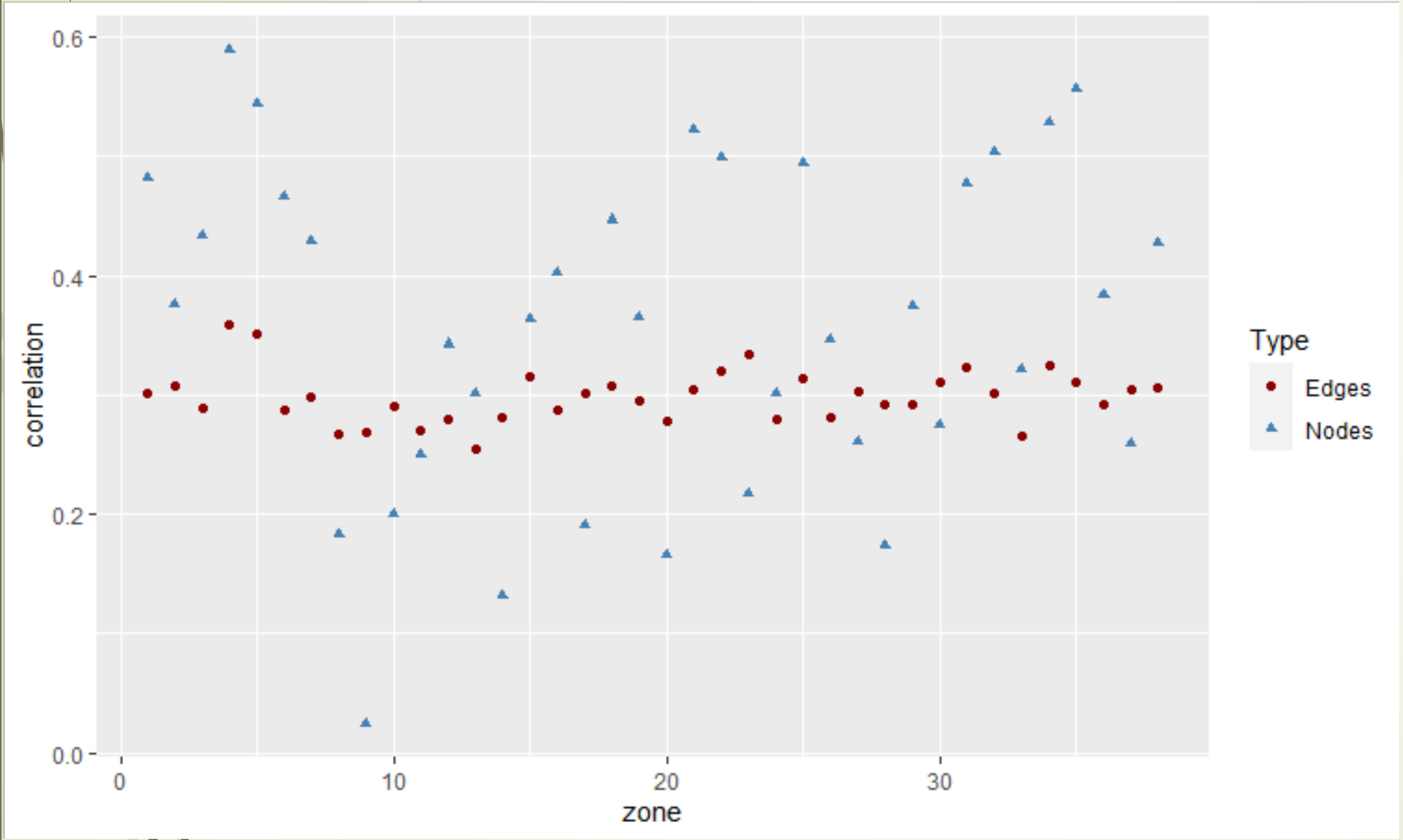
- In particular, focusing on road segments and junctions, the node and edge betweenness appears as key indicators for this context. The node betweenness is a function of the number of shortest paths between pairs of nodes that pass through that node (see Newman, Girvan, 2004):

$$b_i = \sum_{\substack{h,k \in V \\ h \neq k \neq i}} \frac{n_{h,k}(i)}{n_{h,k}}$$

where $n_{h,k}$is the number of shortest paths between h and k and $n_{h,k}(i)$ is the number of shortest paths between h and k that passes through the node i. A similar definition can be provided in case of edges.

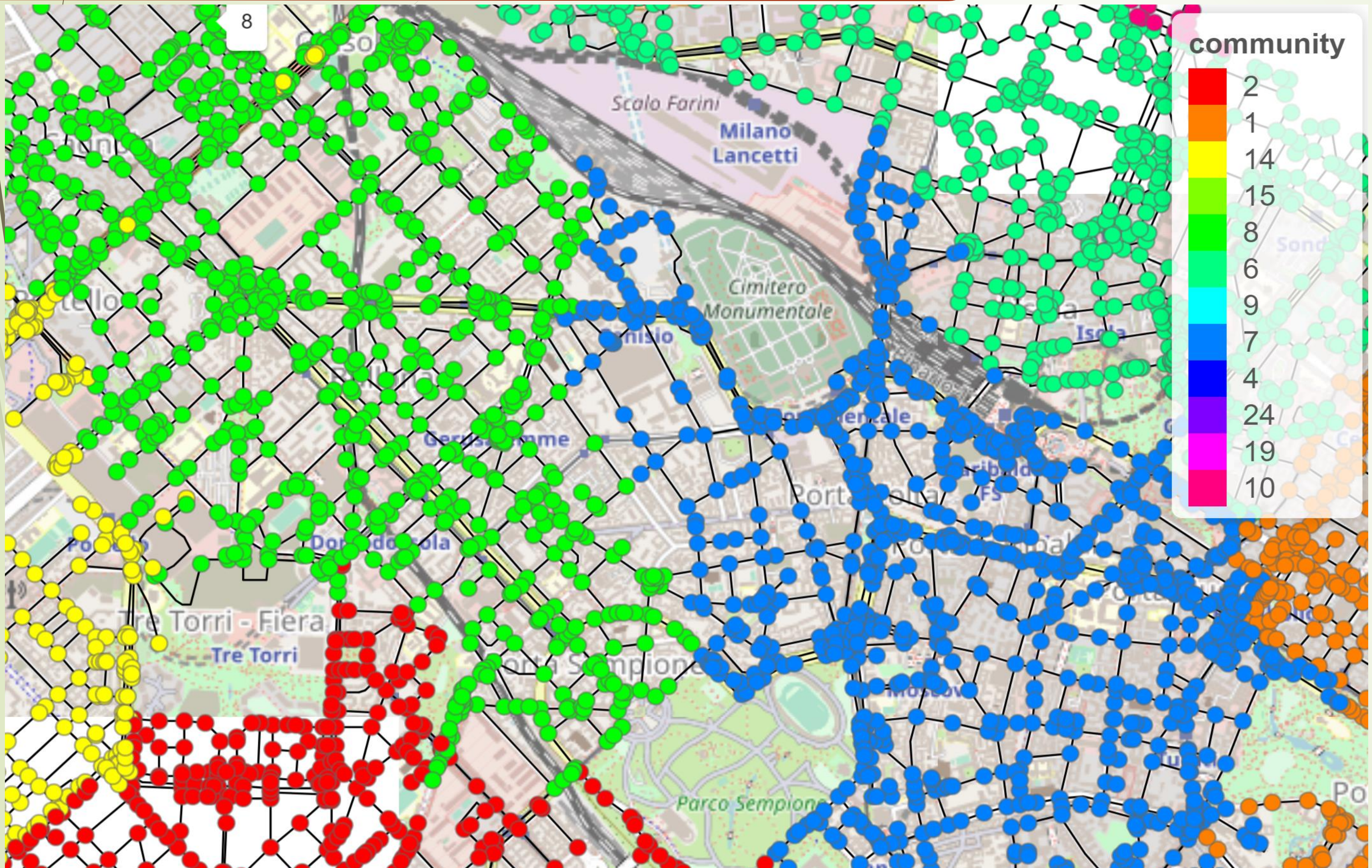
- Since the computation on the whole network G is really time consuming and does not provide significant value added, we considered separately nodes in the sub-graphs G_z based on the splitting of the whole network according to the ZIP codes.
- Since the computation depends on neighbours, the betwenness of nodes in each sub-graph has been estimated considering the network based on $\cup_{h \in N_Z} G_z$ (with N_Z the set that includes G_z and neighbours of G_z, i.e. all the zip codes that touch G_z).

Betweenness vs risk



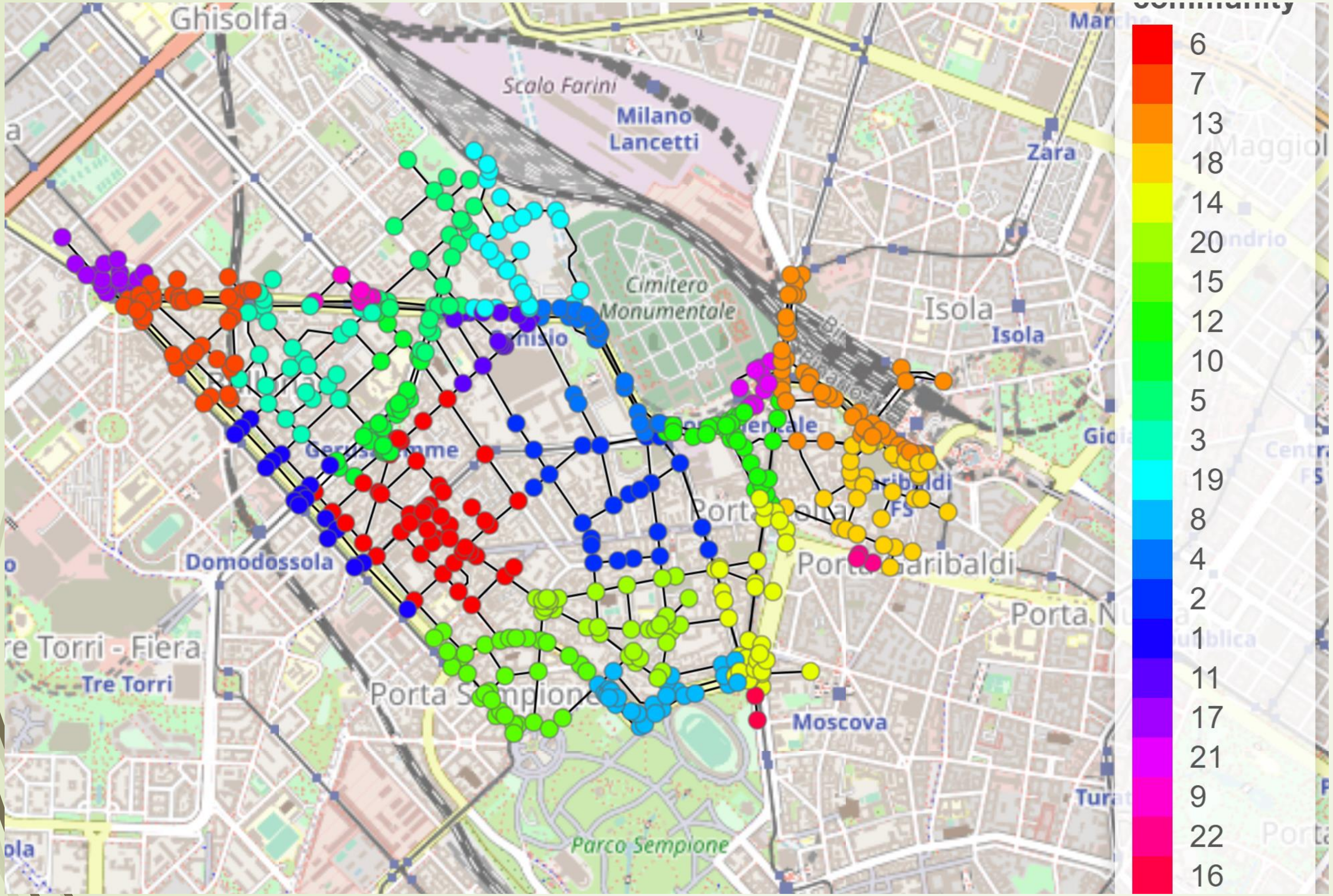
- We display here for each zone (ZIP code of Milan city) ***the rank correlation between the betweenness, based on node or edge, respectively, and the risk assessed***.

Communities



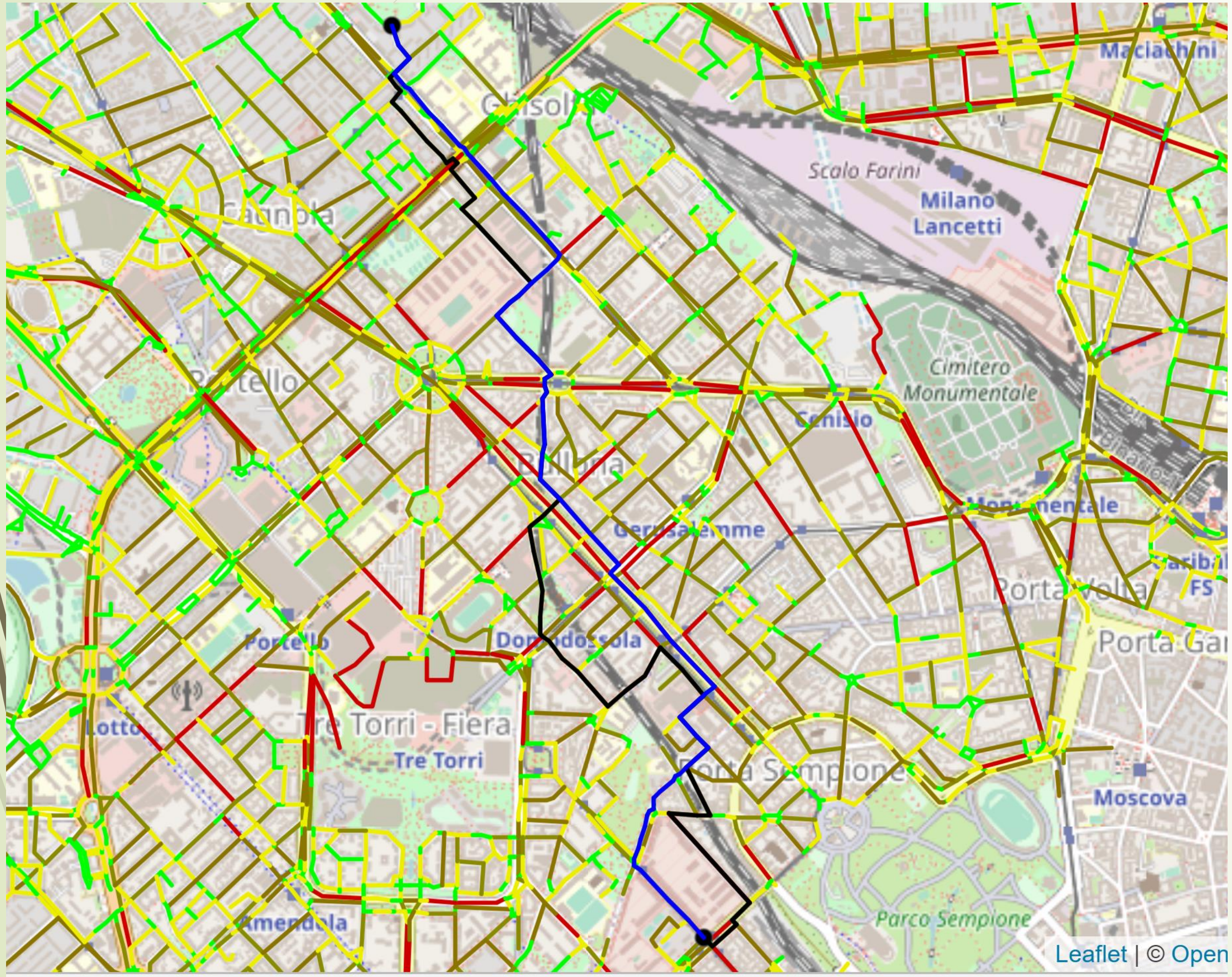
- Considering the weighted graph, **we extract communities** using the Louvain methodology.
- The communities depend on both the edge density and on the weights.
- We find 287 communities considering the whole area of Milan (city and province).
- In the plot communities are ranked according to the average risk

Communities: a higher granularity



- If a higher level of granularity is needed, the methodology could be applied on a subgraph. For instance here a single area (ZIP code) is considered.

Another use of Shortest Path



- We apply the shortest path between two points.
- **In the plot, the shortest path has been applied considering:**
  - **The minimum length (in blue)**
  - **The minimum risk (in black)**

Conclusions and further research

- The proposed approach exploits the use of open-source data to estimate the risk related to where the policyholder drives.
- It is a work in progress and several points are under investigation. In particular, at moment, we are evaluating the possibility of:
  - Improve results using traffic data
  - Compare with other models of spatial analysis to include dependence between segments
  - Using other supervised/unsupervised statistical learning algorithms to estimate the local risk of frequency
  - Evaluate which improvements these results can offer for insurance pricing.

Main references

- Assunção R., Azevedo Costa M., Oliveira Prates M., and Silva e Silva L.G.(2014) Spatial Analysis, in A. Charpentier (2014), *Computational Actuarial Science with R*, Chapman & Hall/CRC press,
- Borgoni, R., Gilardi, A., Zappa, D. (2020), Assessing the Risk of Car Crashes in Road Networks, *Social Indicators Research*
- Boskov M. and Verrall R. J. (1994) Premium Rating at Geographic Area Using Spatial Models. *ASTIN Bulletin*, 24, pp 131-143
- Gilardi,A., Mateu, J., Borgoni, R., Lovelace, R. (2020), Multivariate hierarchical analysis of car crashes data considering a spatial network lattice, Working Paper, ArXiv
- Marshall et al. (2018), Street Network Studies: from Networks to Models and their Representations, Network and Spatial Economics
- Rashmi, R. et al. (2019), Analysis of Road Networks Using the Louvian Community Detection Algorithm, Soft Computing for Problem Solving.
- Tufvesson, O. et al. (2019) Spatial statistical modelling of insurance risk: a spatial epidemiological approach to car insurance, *Scandinavian Actuarial Journal*, 2019:6, 508-522
- Yao J. (2016) Clustering in General Insurance Pricing. In E. Frees, G. Meyers, & R. Derrig (Eds.), *Predictive Modeling Applications in Actuarial Science* (International Series on Actuarial Science, pp. 159-179). Cambridge: Cambridge University Press.
- Wuthrich, M. V., and C. Buser (2019), Data analytics for non-life insurance pricing, g. Available online: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2870308